Московский государственный университет имени М. В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 517 ГРУППЫ

ПОИСК ИНФОРМАТИВНЫХ ПРИЗНАКОВ НА МАГНИТОГРАММНЫХ ИЗОБРАЖЕНИЯХ СОЛНЦА

Выполнил:

студент 5 курса 517 группы

Лаптев Дмитрий Анатольевич

Научный руководитель:

к.ф-м.н.

Ветров Дмитрий Петрович

Москва, 2011

Содержание

Аннотация

В работе рассматривается задача выделения информативных признаков из магнитограммных изображений Солнца для задачи прогнозирования кратковременной солнечной активности.

Построение признаков строится в несколько этапов: сегментация магнитограммы, выделение активных областей на изображении, непосредственный подсчет признаков.

Сегментация проводится с помощью вариационного приближения, параметры марковского случайного поля подбираются автоматически. Процедура подбора параметров направлена на повышение точности результирующего алгоритма на модельной задаче классификации. В качестве алгоритма классификации выбраны решающие леса. Для оценки точности получаемого алгоритма используется скользящий контроль.

Поиск активных областей осуществляется с помощью модифицированного метода ветвей и границ. В работе рассматривается несколько видов функционалов, характеризующих активность области.

На основании рассчитанной области и сегментации, рассчитываются признаки. В работе формализовано понятие "нейтральной линии", которая активно используется в теоретических работах астрономов. Информативность сгенерированного набора признаков доказывается успешной реализацией алгоритма краткосрочной прогнозирования активности.

1 Введение

Практически любая прикладная задача машинного обучения начинается с поиска информативных признаков. Так например, хорошо известны следующие задачи классификации, которые решаются с использованием набора признаков, индивидуального для каждой задачи:

- Медицинская диагностика: требуется поставить диагноз. Признаками являются различные медицинские характеристики, данные анализов и т.д.
- Геологоразведка: определить наличие полезных ископаемых в данной области. Признаками являются результаты зондирования почв.
- Кредитный скоринг: принять решение о выдаче, либо отказе кредита. Признаками являются ответы клиента на анкету.
- Синтез химических соединений: принять или отвернуть гипотезу о наличии свойства получаемого соединения. Признаками являются свойства химических элементов.

Как правило, информативные признаки подбираются на основании мнения экспертов. То есть выделяются признаки из тех, которые используются прямо или косвенно экспертами при решении подобных задач.

Нами рассматривается задача выделения информативных признаков для задачи краткосрочного прогнозирования солнечной активности (подробное описание задачи и ее решение смотрите в [1]). В рассматриваемой задаче экспертные признаки оказываются слабо формализованными, что усложняет задачу построения автоматической системы генерации признаков.

В следующих главах рассматривается постановка задачи и обзор релевантной литературы, далее вводится подробное описание входных данных, их формат, источники, особенности. В главе "Схема выделения информативных признаков" вводится общая схема алгоритма, в следующих главах раскрывается подробнее каждый из этапов алгоритма: сегментация, поиск активных областей, подсчет признаков. В последних главах представлена схема автоматического подбора параметров марковских случайных полей.

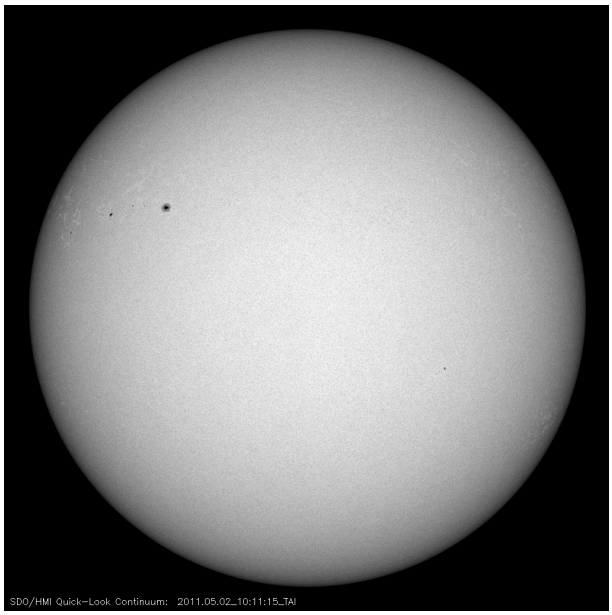

Рис. 1: Солнце в видимом диапазоне

2 Постановка задачи

Задача выделения информативных признаков возникла как один из основных шагов для решения более глобальной задачи. Институтом Космических Исследований (ИКИ, [2]) была поставлена задача краткосрочного прогнозирования солнечной активности. На данный момент активность Солнца отслеживается группой экспертов, которые следуя эвристическим алгоритмам, определяют, насколько активно будет вести себя Солнце в ближайшее время. Автоматических систем, способных заменить группу экспертов, на данный момент не создано.

С точки зрения формальной постановки задачи, солнечная активность характеризуется солнечными вспышками, которые, в свою очередь, связаны с такими явлениями, как солнечные пятна. Кроме того, существует много косвенных признаков солнечной активности, такие, как, например, показатели магнитного поля и уровень радиации в околоземном пространстве. Таким образом, для задачи краткосрочного прогнозирования солнечной активности могут быть использованы различные данные, релевантные задаче: снимки солнечной поверхности со спутника в видимом диапазоне (иллюстрация 1), магнитограммы, корональные снимки Солнца (иллюстрация 2), измерения полей в околоземном пространстве (иллюстрация 3), положение планет солнечной системы и т.д.

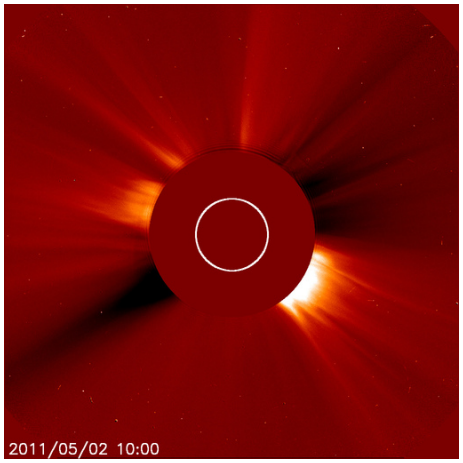

Рис. 2: Корональный снимок Солнца

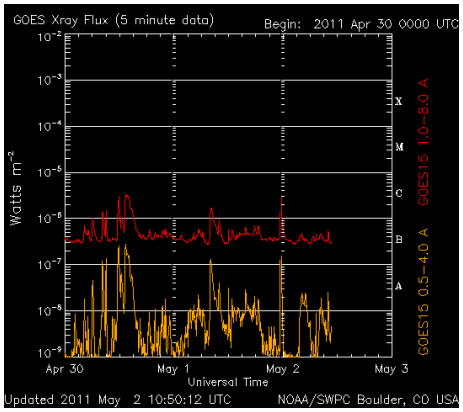

Рис. 3: Измерение радиации в околоземном пространстве

Для выбора наиболее информативных и удобных для анализа инструментов, был проведен ряд совещаний с экспертной группой Института Космических Исследований. По результатам обсуждения задачи было выявлено, что магнитограммы являются основным инструментом для исследования солнечной активности экспертами, так как отражают первоисточник активности солнца — структуру и силу магнитных полей на поверхности Солнца. Таким образом было принято решение анализировать в первую очередь магнитограммы, выделяя из них информативные признаки.

Далее в работе речь пойдет о формировании информативных признаков из магнитограммных изображений Солнца.

Формализованная постановка задачи звучит следующим образом: необходимо построить алгоритм и программный модуль, который выделял бы информативные признаки на основании последовательности магнитограмм во времени. Информативные признаки должны строиться полностью автоматизировано и формализовано, чтобы исключить присущую человеку субъективность. Информативные признаки должны быть построены на основании признаков, выделенных экспертами, таким образом, чтобы максимизировать точность прогноза в задаче краткосрочного прогнозирования солнечной активности.

3 Обзор аналогов

Задача выделения информативных признаков из изображений Солнца не является новой. Даже в то время, когда еще не было специальной аппаратуры для проведения сложных измерений, предпринимались активные попытки формализовать формы солнечных пятен в видимом диапазоне. Примером такой формализации, предложенной в 1966 году МакИнтошом в работе [3], является классификация солнечных пятен на основе видимого диапазона.

Классификация МакИнтоша разделяла группы пятен на солнце на несколько классов, каждый из которых характеризовал активность группы и, соответственно, гипотетическую возможность того, что данная группа пятен станет причиной вспышки на Солнце. Данная классификация до сих пор является мировым стандартом в

описании данных видимого диапазона, аккумулируя в себе множество характеристик группы солнечных пятен, таких как:

- количество пятен в группе,
- размер пятен,
- размер тени и полутени,
- форму пятен.

В работе [4] описывается алгоритм автоматической классификации пятен по МакИнтошу без участия человека по снимкам солнечной поверхности со спутника SOHO.

Одна из первых работ, в которой предпринимается попытка формализовать признаки, выделяемые из магнитограммных изображений Солнца, является работа [5]. Она посвящена классификации магнитных структур, возникающих на Солнце. Авторы, по аналогии с классификацией МакИнтоша, учитывают различные факторы, которые характеризуют активность области:

- являются ли области биполярными,
- размер областей,
- размер положительного и отрицательного полей,
- форма областей,
- взаимопроникновение областей положительного и отрицательного магнитного полей.

В статье [6] предложен метод прогнозирования солнечной активности, который основан на нескольких признаках, в том числе на признаках, получаемых из магнетограмм:

- класс по МакИнтошу,
- класс магитного поля,
- площадь группы пятен,

- изменение интенсивности излучения в диапазоне 10.7 см.,

Значения всех признаков учитываются с весами, которые были подобраны экспертами. Недостатком данного алгоритма является небольшая точность при условиях сбалансированности выборки.

Один из важнейших признаков, озвученных экспертами Института Космических Исследований, является "нейтральная линия" — слабо формализованное понятие, которым обозначается линия между областями положительного и отрицательного поля. Первые попытки формализации этого признака приводятся в [7], и позднее, уже в контексте прогностического признака для задачи прогнозирования активности, в работе [8]. Тем не менее, в этих работах не было фиксировано четкого алгоритма, который бы описывал построение этой линии, во многом допуская различную интерпретируемость. Алгоритмы, изложенные в данных работах не были реализованы в виде прогаммного модуля, скорее они были направлены на доказательство информативности данного признака.

В 2011 году была представлена первая работа [9], в которой нейтральная линия была использована для построения программы, прогнозирующей солнечную активность. В реализации авторов работы использовался единственный признак: величина потока, проходящего через нейтральную линию, но само понятие нейтральной линии опять же не было формализовано.

Уже на этапе обзора аналогов мы неоднократно упоминали без формализации такие понятия, как "область положительного магнитоного поля", "область отрицательного магнитного поля", "нейтральная область". Для формализации этих параметров в дальнейшем изложении работы используется алгоритм сегментации изображений. Задача сегментации является задачей дискретной оптимизации, сформулированной в работе [11]. Первые методы ее решения опубликованы в книге [10] .

Одним из наиболее распространенных алгоритмов классификации является алгоритм разреза графов, предложенный в статье [12]. Активное использование алгоритмов нахождения максимального потока в графе через нахождение разреза графов для задачи сегментации началось со статьи [15].

Основным недостатком данного метода для нашей задачи является тот факт, что в явном виде он неприменим к задаче сегментации на число классов, большее

двух. Для решения подобных задач существует несколько приближенных методов, получивших наибольшее распространение.

Первый из них, альфа-расширение, был предложен в статье [16] и основан на итеративном применении метода разреза графов. Другие, такие как Tree Reweighted Message Passing (TRW), изложенный в [13], [14], основаны на разбиение задачи дискретной оптимизации на циклическом графе на подзадачи на деревьях, для которых может быть найдено точное решение.

Другим методом, который используется в данной работе, является вариационное приближение, предложенное в работе [22]. Данный метод имеет функционал, аналогичный по сравнению с методом альфа-расширения, но позволяет учесть также некоторые глобальные ограничения ([23], [24]) на такие характеристики как, например, общую площадь положительного и отрицательного потоков, их суммы. Данные ограничения имеют физический смысл, так как, например, в большинстве активных областей отрицательный поток равен положительному.

Подробнее про метод вариационного приближения рассказано в разделе "Сегментация".

4 Описание входных данных

Входными данными, на основании которых необходимо было построить алгоритм и выделять информативные признаки, является магнитограммное изображение Солнца, или MDI магнитограмма Солнца (MDI — Michelson Doppler Imager). Данные изображения снимаются с помощью специального оборудования, смотрите подробнее [17], [18]. Изображение MDI магнитограммы отображает магнитные полюса в фотосфере Солнца, в черно белом представлении, Противоположные магнитные полюса отображены боле или менее светлыми градациями серого цвета. На рисунке 4 приведен пример магнитограммного изображения Солнца. Размер представленного изображения в оригинале равен $1024 * 1024$. На рисунке 5 представлен фрагмент магнитограммы, где крупно показана область, на которой мы, посмотрев на Солнце через закопченное стекло, увидели бы солнечное пятно. Белые области здесь соответ-

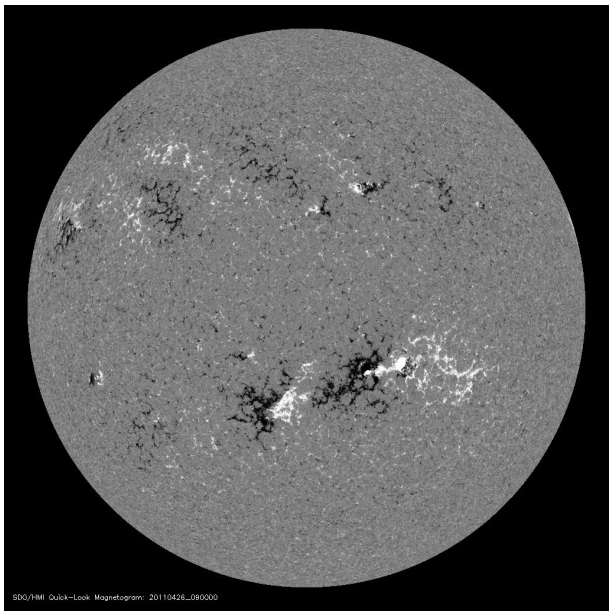

Рис. 4: Пример магнитограммного изображения Солнца

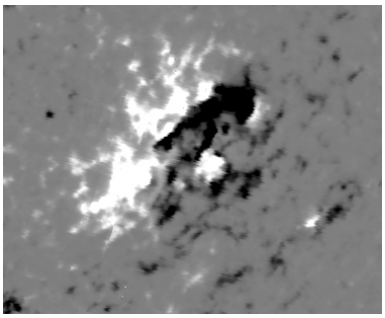

Рис. 5: Солнечное пятно на магнитограмме

ствуют сильному отрицательному полю, черные - сильному положительному, серый — области с низким магнитным полем (близким к нулю).

Данные изображения получаются уже после обработки, до обработки они представляют из себя данные в специальном формате FITS ([19]). FITS — Гибкая Система Передачи Изображений (англ. Flexible Image Transport System), цифровой формат файлов используемый в науке для хранения, передачи и редактирования изображений и их метаданных. Чаще всего FITS используется в астрономии. В отличие от других форматов изображений, FITS разработан специально для научных данных и потому включает в себя метаданные, описывающие информацию о фотометрической и пространственной калибровке, вместе с метаданными исходного изображения.

Говоря более конкретно об магнитограммных изображениях Солнца, в данных формата FITS хранится информация об абсолютном значении магнитного поля в каждой точке сетки поверхности Солнца (сетка задается исходя из разрешающей способности приборов, в нашем случае, как уже было сказано, $1024 * 1024$), а также дополнительная информация о калибровке приборов, ориентации спутника, времени и т.д.

Абсолютные значения магнитного поля на солнечной поверхности измеряются в Гауссах, в областях на магнитограмме, где мы видим "черные пятна", значения поля достигают, иногда, нескольких тысяч Гауссов, в областях, где наблюдаются "белые пятная" — нескольких тысяч со знаком минус. Серые области соответствуют значению магнитного поля примерно до двухсот Гауссов по модулю.

По рекомендации специалистов Института Космических Исследований, был определен ресурс, с которого должны поступать данные для алгоритма: спутник-обсерватория SOHO [20]. SOHO (англ. Solar and Heliospheric Observatory, код обсерватории «249») — космический аппарат для наблюдения за Солнцем. Совместный проект ЕКА и НАСА. Был запущен 2 декабря 1995 и приступил к работе в мае 1996. Имеет на борту 12 инструментов, позволяющих получать изображения и/или измерять потоки излучения Солнца.

На сайте НАСА (http://sohowww.nascom.nasa.gov/) для скачивания доступны снимки солнечной поверхности, в их числе магнитограммные изображения. Срок, за которых хранятся данные на сайте — с момента начала работы спутника, то есть более, чем за 10 лет. Периодичность, с которой происходит съемка солнечной поверхности — несколько раз в день, до восьми снимков в сутки, то есть каждые три часа.

Кроме этого, мы рассматривали только данные о магнитограммах, которые были привязаны по времени к солнечным вспышкам (данные о солнечных вспышках собраны в отдельной базе данных, [9]). В итоге в рассмотрение попадает более 10000 магнитограмм.

5 Схема выделения информативных признаков

Совместно с экспертами были выделены несколько неформальных признаков, которые используются при ручном анализе солнечной активности:

- размер областей с сильным положительным и отрицательным полями,
- интегральные характеристики поля,
- сложность структуры поля,
- сложность взаимодействия полей,
- динамика развития полей,
- ряд других признаков.

Стоит особо подчеркнуть, что данные признаки не являются формальными, правила их использования, сформулированные экспертами, допускают довольно широкую интерпретацию.

Некоторые из этих признаков базируются на таких понятиях, как "область с сильным положительным полем" и "область с сильным отрицательным полем". Эти понятия не формализованы точно, и их формализация допускает различные варианты. Но ясно, что для подсчета признаков, требуется каким-то образом определять, какие области на Солнце являются активными, какие области можно назвать областями с сильным положительным или отрицательным полем. Следовательно, необходимо было выделить данный этап и реализовать алгоритм, который бы с формальной точки зрения отделял упомянутые выше понятия.

Кроме того, следует сказать, что в процессе подсчета признаков, необходимо привязываться не к солнечному диску в целом, а только к некоторой его активной области, которая и является источников солнечных вспышек. То есть, необходимо локализовать регион Солнца, в которой находится область повышенной активности. Это необходимо делать также по той причине, что на Солнце может быть несколько активных областей, но чаще всего только одна из них является причиной солнечных вспышек и представляет интерес для исследования. Таким образом, второй этап,

который возникает на пути подсчета признаков — локализация рассматриваемой области.

На основании локализованных областей повышенной активности уже можно вычислять некоторые характеристики текущей конфигурации магнитных потоков.

Резюмируя все вышесказанное, схема генерации признаков состоит из трех этапов:

- выделение областей повышенной активности на магнитограммных изображениях Солнца,
- локализация регионов повышенной активности,
- непосредственный подсчет информативных признаков.

Выделение областей повышенной активности было реализовано посредством сегментации исходного массива на три сегмента.

Локализация регионов повышенной активности проводилась путем выбора области, на которой достигается максимум заданного функционала (решение осуществлялось при помощи метода ветвей и границ).

Непосредственный подсчет включает в себя набор алгоритмов для подсчета некоторого фиксированного набора признаков. Набор алгоритмов максимально соответствует экспертным и является четкой формализацией экспертных признаков.

Каждый из этих шагов будет подробнее рассмотрен далее.

6 Сегментация

Первый этап выделение областей с сильным положительным и отрицательным полем, может быть переформулирован в терминах известного вида задач классификации: задачи сегментации изображений.

Дано изображение, пиксели которого требуется отнести к одному из трех классов:

- пиксели, которые относятся к сильному положительному магнитному полю,
- пиксели, которые относятся к сильному отрицательному магнитному полю,

- пиксели, которые не относятся ни к одному из первых двух классов, "нейтральные пиксели".

Будем в дальнейшем обозначать пиксель изображения i, значение поля в этом пикселе I_i, соответственно, наблюдаемое изображение, массив всех данных о значении поля — I.

Сегментация может быть проведена разными способами. Одним из простейших способов, которым можно воспользоваться для сегментации в случае нашей задачи, является, например, отсечением по порогу:

- пиксель i, в котором поле I_i больше заданного порога C^+ относятся к классу "сильная положительная область",
- если I_i меньше заданного порога C^-, то пиксель относится к классу "сильная отрицательная область",
- все остальные пиксели относятся к классу "нейтральная область".

Недостатком данного метода является то, что получившаяся сегментация является крайне нестабильной: соседние пиксели очень часто относятся к разным классам. С этими недостатками позволяет справиться использование следующих методов сегментации.

6.1 Марковские случаные поля

Введем несколько понятий, которые необходимы нам для формальной постановки задачи.

N — общее число пикселей (применительно к нашей задаче $N = 1024 * 1024 = 1048576$),

Z_i — номер класса, к которому относится пиксель i, соответственно Z — общая конфигурация по всем пикселям,

$\mathcal{E}$ — область связности: пиксель j находятся в отношении связности с пикселем i, если $j \in \mathcal{E}(i)$,

$\varphi_i(Z_i)$ — унарный множитель, характеризует то, насколько метка класса Z_i "хорошо" подходит для пикселя i, зависит от значения наблюдаемой переменной в пикселе i,

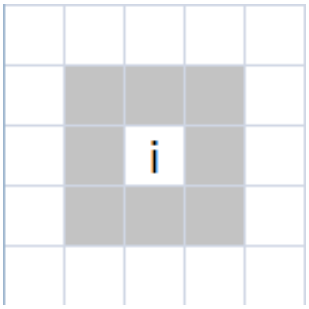

Рис. 6: Пример системы соседства

$\phi(Z_i, Z_j)$ — бинарный множитель, который называется обобщенной моделью Поттса [15], характеризует то, насколько метка класса Z_i подходит для пикселя i, в условиях того, что для пикселя j была выбрана метка Z_j.

В этих терминах можно сформулировать следующую задачу дискретной оптимизации:

$$p(I|Z) \propto \prod_{i=1}^{N} \varphi_i(Z_i) \prod_{j \in \mathcal{E}(i)} \phi(Z_i, Z_j) \to \max_{Z}. \tag{1}$$

Простая интерпретация данной задачи может быть сформулирована следующим образом: "требуется приписать пикселям метки классов таким образом, чтобы, с одной стороны, учесть характеристики поля в данной точке, а с другой стороны — учесть метки классов в соседних пикселях".

Общепринятой практикой является использовать в качестве бинарного множителя функцию, которая равна единице, если метки классов совпадают, и меньше единицы, если не совпадают:

$$\phi(Z_i, Z_j) = e^{C_{pair}[Z_i \neq Z_j]}. \tag{2}$$

Здесь квадратные скобки "[...]" — скобки Иверсона, данное выражение равно 1, если значение в скобках истинно и равно 0 иначе.

В качестве системы соседства было решено использовать восьмисвязные области: пиксель j находится в отношении соседства с пикселем i, если они физически являются соседними пикселями (см. рисунок 6).

Остается вопрос, каким образом лучше задать унарный потенциал. После ряда экспериментов в качестве унарного множителя был выделен класс функций, которые работают стабильнее всего:

$\varphi_i(1) = e^{-C_1\sqrt{|2000-I_i|}}$,
$\varphi_i(2) = e^{-C_1\sqrt{|2000+I_i|}}$,
$\varphi_i(3) = e^{-C_2|I_i|}$.

Таким образом, мы получили модель, которая зависит от трех параметров:

- $_1$ — ширина функционала, "отвечающего" за активные области,
- $_2$ — ширина функционала, "отвечающего" за нейтральные области,
- $_{pair}$ — константа, показывающая, насколько сильно необходимо учитывать систему соседства при сегментации.

Может быть показано, что задача вида 1 является NP-трудной, если число меток классов больше двух. Как уже было сказано в разделе "Обзор аналогов", существует несколько зарекомендовавших себя методов приближенной максимизации заданного функционала, в дальнейшей работе речь пойдет об одном из них: вариационном приближении.

6.2 Вариационное приближение

Вариационное приближение — один из методов приближенного решения задач дискретной оптимизации, идея которого заключается в том, чтобы приблизить изначальное распределение факторизованным распределением, имеющим простой вид:

$$p(Z|I) \approx q(Z) = \prod_{i=1}^{N} q_i(Z_i), \tag{3}$$

Логично предположить, что распределение $q(I)$ следует выбирать таким образом, чтобы оно как можно более точно приближало изначальное распределение $p(I)$. Достигнуть этого можно, например, минимизацией дивергенцией Кульбака-Лейблера между распределениями $q(Z)$ и $p(Z|I)$ [25].

$$\mathrm{KL}(q||p) = -\int q(Z) \log \frac{p(Z|I)}{q(Z)} dZ \to \min_{q(Z)}.$$

Подставим факторизованное приближение в выражение для дивергенции $\mathrm{KL}(q(Z)||p(Z))$:

$$\mathrm{KL}(q(Z)||p(Z)) = -\int \prod_i q_i \left(\log p(Z) - \sum_i \log q_i \right) dZ =$$

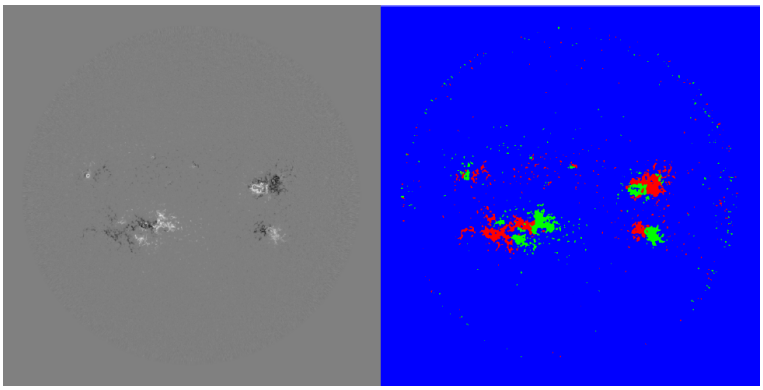

Рис. 7: Исходное изображение и результаты сегментации

$$= -\int q_j \left(\int \log p(Z) \prod_{i \neq j} q_i dZ_i \right) dZ_j + \int q_j \log q_j dZ_j + C$$

Обозначив $\log \hat{p}(Z_j) = \mathrm{E}_{i\ nej} \log p(Z) + C$, тогда

$$\mathrm{KL}(q||p) = \mathrm{KL}(q_j||\hat{p}) + C$$

Оптимальное распределение $q_j^*(Z_j) = \hat{p}(Z_j)$, то есть:

$$\log q_j^*(Z_j) = \mathrm{E}_{i\ nej} \log p(Z) + C$$

$$q_j^*(Z_j) = \frac{\exp(\mathrm{E}_{i\ nej} \log p(Z)}{\int \exp(\mathrm{E}_{i\ nej} \log p(Z) dZ_j}$$

Это и есть основная формула вариационного приближения. Подставляя конкретный вид распределения, получим следующее соотношение:

$$q_i(Z_i) = \frac{1}{C} \exp \left(\log(\varphi_i(Z_i)) - C_{pair} \sum_{t \in \mathcal{E}(i)} \sum_{j \neq i} q_j(Z_j) \right)$$

Нахождение решения данного уравнения в явном виде невозможно, но легко построить итерационный процесс, который будет за несколько итераций сходится к решению:

$$q_i^{new}(Z_i) = \frac{1}{C} \exp \left(\log(\varphi_i(Z_i)) - C_{pair} \sum_{t \in \mathcal{E}(i)} \sum_{j \neq i} q_j^{old}(Z_j) \right)$$

Пример исходного изображения и результатов сегментации для параметров $C_{pair} = 20$, $C_1 = C_2 = 1$ представлены на рисунках 7 и 8.

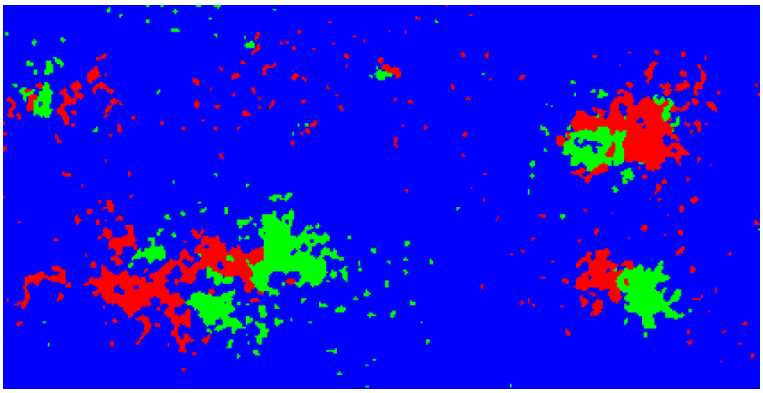

Рис. 8: Результаты сегментации

7 Поиск активных областей на изображении

Для локализации интересующих нас областей на магнитограммных изображении Солнца необходимо выбрать некоторый критерий того, что данная область является наиболее активной. Будем искать активный регион как прямоугольную область, ограничивающую сильное отрицательно и положительное поля.

Будем использовать метод, предложенный в статье [26], который использует известный метод ветвей и границ для нахождения максимума функционала на прямоугольнике.

Обозначив прямоугольник R, можно выбрать функционал следующим образом:

$$F(R) = \alpha \sum_{i \in R} A_i - \sum_{i \in R} B_i \to \max_R \tag{4}$$

Здесь $A_i = q_i(1) + q_i(2)$ and $B_i = q_i(3)$.

Прямоугольник R однозначно задается координатами верхней границы, нижней, левой и правой: (t, b, l, r). Чтобы описывать не один прямоугольник, а множество, можно заменить число-координату на множество. $\mathbf{R}$ зададим четырьмя интервалами $[T, B, L, R]$, где $T = [t_{low}, t_{high}]$ и так далее.

Число прямоугольников конечно, следовательно, если на каждом шаге алгоритме сокращать размер множества, мы дойдем до оптимального прямоугольника за конечное число шагов. Но перебирать все возможные прямоугольники — очень ресурсоемкая задача, не выполнимая за приемлемое время на современных компьютерах, поэтому было решено использовать следующую оптимизацию.

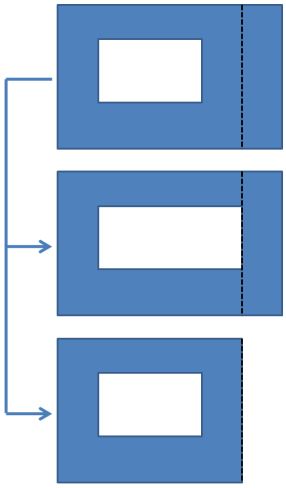

Рис. 9: Схема процедуры ветвления

Необходимо найти функционал $\hat{F}(\mathbf{R})$, который является верхней границей для функционала $F(R)$ такой, что:

- $\hat{F}(\mathbf{R}) \geq F(R), \forall R \in \mathbf{R}$,
- $\hat{F}(\mathbf{R}) = F(R)$, если R – единственный элемент множества $\mathbf{R}$,

Введем очередь с приоритетами, на первом шаге инициализируем очередь единственным элементом, который совпадает со всем множеством прямоугольников. Далее будем последовательно применять метод ветвей и границ:

- Выберем из очереди первый элемент $\mathbf{R}$ (с наибольшим значением $\hat{F}(\mathbf{R})$).
- Произведем процедуру ветвления: разобьем множество $\mathbf{R}$ на два подмножества $\mathbf{R}_1$ и $\mathbf{R}_2$, составляющие в сумме все множество.
- Для каждого из новых множеств прямоугольников подсчитаем значение функционала $\hat{F}(\mathbf{R}_1)$ и $\hat{F}(\mathbf{R}_2)$ и поместим в очередь.
- Если на каком-то шаге алгоритма первый элемент очереди будет состоять из одного прямоугольника, то найден максимум функционала F.

Схематично процедура ветвления представлена на рисунке 9.

Функция, являющаяся границей сверху, может быть выписана, например, следующим образом:

$$\hat{F}(\mathbf{R}) = \alpha \sum_{i \in R_{small}} A_i - \sum_{i \in R_{big}} B_i$$

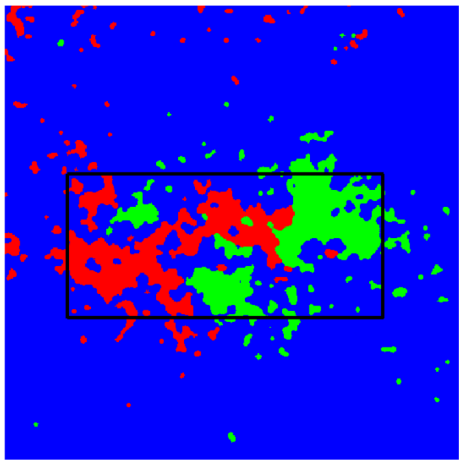

Рис. 10: Результат поиска ограничивающего прямоугольника

Где R_{small} — наименьший прямоугольник в множестве $\mathbf{R}$, он задается координатами $(t_{high}, b_{low}, l_{high}, r_{low})$, R_{big} — напротив, наибольший прямоугольник $(t_{low}, b_{high}, l_{low}, r_{high})$.

Результат данной процедуры при значении параметра $\alpha = 4$ представлен на рисунке 10.

Результат поиска ограничивающего прямоугольника, полученного в результате проделанной процедуры, оказался не совсем таким, каким его хотелось бы видеть экспертам. так например, есть некоторые активные области, которые не попали в прямоугольник. Бороться с этим эффектом позволяет выбор другого функционала, значение которого также зависит от того, сколько пикселей активной области лежит на границе:

$$F(R) = \alpha \sum_{i \in R} A_i - \sum_{i \in R} B_i + \beta \sqrt{Area(R)} \sum_{i \in border\ of\ R} B_i.$$

Интересным вопросом при таком выборе функционала является вопрос выбора верхней границы $\hat{F}(\mathbf{R})$. Многие из них вариантов, которые были опробованы, либо не удовлетворяли некоторым условиям, наложенным на функционал, либо требовали значительного времени на вычисление. В конечном итоге выбор остановился на функционале следующего вида:

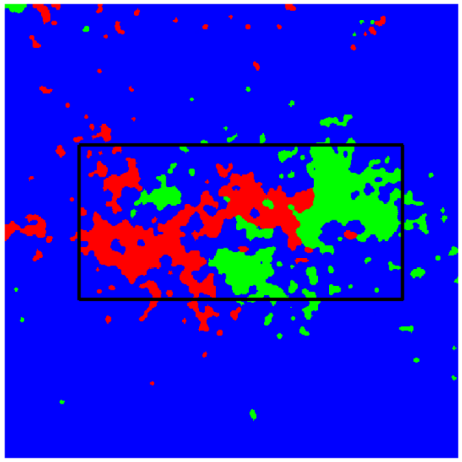

Рис. 11: Ограничивающий прямоугольник с учетом граничной информации

$$\hat{F}(\mathbf{R}) = \alpha \sum_{i \in R_{small}} A_i - \sum_{i \in R_{big}} B_i + \beta \sqrt{Area(R_{big})} (\max_{R \in \mathbf{R}} \sum_{i \in left\ border\ of\ R} B_i +$$

$$\max_{R \in \mathbf{R}} \sum_{i \in top\ border\ of\ R} B_i + \max_{R \in \mathbf{R}} \sum_{i \in right\ border\ of\ R} B_i + \max_{R \in \mathbf{R}} \sum_{i \in bottom\ border\ of\ R} B_i). \quad (5)$$

Данный функционал также может быть вычислен очень эффективно. Время работы алгоритма составляет менее секунды на изображении $1024 * 1024$.

Результат работы для того же участка изображения с параметрами $\alpha = 4, \beta = 0.05$ представлен на рисунке 11. Везде далее считается, что эти параметры фиксированы, их выбор обусловлен экспертными оценками результатов.

8 Подсчет признаков

После того, как были выделены области повышенной солнечной активности, можно переходить к непосредственному вычислению информативных признаков.

Одними из наиболее информативных признаков по мнению экспертов являются характеристики так называемой "нейтральной линии", линии, разделяющей области с сильным положительным и отрицательным полем. С помощью этой линии можно понять, насколько сильно взаимодействуют отрицательные и положительные поля, насколько сложна структура полей.

8.1 Нейтральная линия

Формально понятие нейтральной линии не вводится ни в одной из аналогичных работ, мы использовали следующее определение: *Нейтральная линия — линия между областями повышенной активности, но разной полярности, которые были получены в процессе сегментации.*

Таким образом, введенная формализация будет базироваться на результатах сегментации изображения Солнца. Дальнейшая формализация будет проведена двумя способами: введем алгоритм, строящий простую линию, и линию, которая будет устойчива к шумам.

8.1.1 Простой алгоритм нейтральной линии

Прямое применение определения, введенного выше, предполагает нахождение всех пикселей, которые имеют соседей как из класса положительного магнитного поля, так и из класса отрицательного магнитного поля. Система соседства может быть выбрана различными способами, в данной работе используется восьмисвязная область. Эффективный алгоритм, реализующий нахождение всех таких пикселей изложен ниже:

1. Создать новое бинарное изображение A, где все пиксели, принадлежащие к положительному классу, отмечены 1, все остальные — 0.

2. Произвести морфологическое расширение изображения A, чтобы получить увеличенную область.

3. Создать новое бинарное изображение B, где все пиксели, принадлежащие к отрицательному классу, отмечены 1, все остальные — 0.

4. Произвести морфологическое расширение изображения B, чтобы получить увеличенную область.

5. Построить пересечение изображений A и B: $r = A * B$.

6. Получившееся изображение состоит из линии между классами, то есть нейтральной линии.

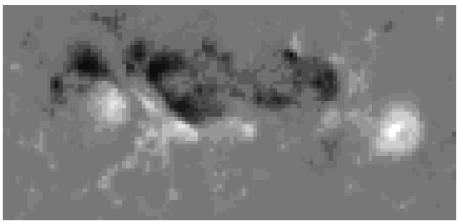

Рис. 12: Исходное изображение

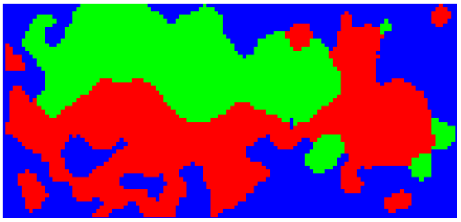

Рис. 13: Сегментация исходного изображени

Результат работы такого алгоритма, примененный к области, изображенной на рисунках 12, 13 представлен на рисунке 14.

8.1.2 Устойчивая нейтральная линия

Можно заметить, что нейтральная линия, полученная в ходе алгоритма, изложенного выше, содержит множество отдельных кусков, которые отделяют совсем небольшие области. Такие области в процессе эволюции пятна могут как появляться, так и исчезать, внося коррективы в нейтральную линию, зашумляя ее. Второй алгоритм нейтральной линии, который было принято решение использовать, как раз и направлен на то, чтобы отсекать небольшие области, делая линию более устойчивой:

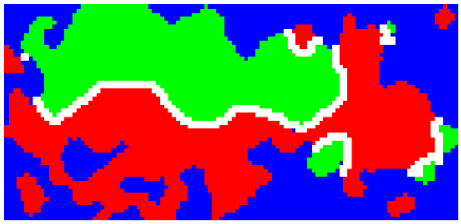

Рис. 14: Простой алгоритм нейтральной линии

Рис. 15: Изображение C

1. Создать новое бинарное изображение, в котором пиксели, принадлежащие либо положительной области, либо отрицательной, отмечены единицей, остальные — нулем.

2. Для каждого связного региона на получившемся изображении:

   (a) Создать новое бинарное изображение A, где все пиксели, принадлежащие к положительному классу, отмечены 1, все остальные — 0.

   (b) Произвести морфологическое сжатие изображения A.

   (c) Провести аналогичную процедуру для отрицательного класса (бинарное изображение B).

   (d) Составить новое изображение $C = A + B$ (см. рисунок 15).

   (e) Убрать из рассмотрения связные области, площадь которых менее одного процента от общей площади всех областей (см. рисунок 16).

   (f) Для каждой из оставшихся областей C_i:

      i. пересечь область C_i и изображение A: $r_A = C_i * A$,

      ii. пересечь область C_i и изображение B: $r_B = C_i * B$,

      iii. пересечь полученные изображения: $r = r_A * r_B$,

      iv. отфильтровать части изображения, длина которых менее, чем $0.15\sqrt{Area_C}$.

3. Результирующая нейтральная линия является объединением всех линий r для каждой связной области (см. рисунок 17).

Некоторые этапы построения нейтральной линии показаны на рисунках 15 — 17. На рисунке 18 приведены для сравнения две нейтральной линии, построенные различными алгоритмами, белый отмечена простая нейтральная линия, черной — устойчивая к шуму.

Рис. 16: Из рассмотрения убраны маленькие области

Рис. 17: Получившаяся нейтральная линия

Рис. 18: Два вида нейтральной линии

8.2 Итоговые признаки

Совместно с экспертами были выделены следующие признаки из магнеготрамм, наиболее полно описывающие текущую активность Солнца:

- Координаты центра выделенного ограничивающего прямоугольника — чем ближе к экватору Солнца находится пятно, тем больше вероятность того, что оно будет являться причиной вспышки, которая окажет влияние на околоземное пространство.
- Число отдельных групп пятен на солнечном диске — характеризует интегральные характеристики солнечной активности.
- Площадь отрицательной и положительной области в ограничивающем прямоугольнике.
- Положительный и отрицательный потоки — сумма по всем пикселям, принадлежащим к положительной (соответственно отрицательной) области.
- Максимальное и минимальное значение магнитного поля в ограничивающем прямоугольнике.
- Длина нейтральной линии — описывает, насколько близки области положительного и отрицательного магнитного поля.
- Площадь изображения, которое получается из нейтральной линии, если провести морфологическое расширение на пять пикселей — чем меньше отношение данного признака и длины нейтральной линии, тем сложнее ее форма взаимодействия полей.
- Сумма потоков в области из предыдущего признака — описывает поток, который проходит через нейтральную линию.
- Сумма положительных потоков в области из предыдущего признака.
- Скорость изменения вышеперечисленных признаков во времени.

8.3 Оценка информативности набора признаков

Чтобы понять, насколько подходят для решения той или иной задачи сгенерированный набор признаков, необходимо запустить некоторый алгоритм машинного обучения на этой задаче, которому на вход будут подаваться сгенерированные признаки. В зависимости от точности получившегося алгоритма можно судить об информативности набора признаков.

Информативность признаков проверяется на одной из подзадач, которая является упрощением задачи прогнозирования солнечной активности. Формулируется она следующим образом.

Пусть нам заданы два порога:

- $C_{time} = 2$ дня — порог по времени до вспышки: экспертов интересуют вспышки, которые случатся в течении времени, меньшего C_{time},
- $C_{magnitude} = M5$ — порог на силу вспышки: экспертов интересуют вспышки, которые имеют силу больше, чем $C_{magnitude}$.

Соответственно вспышки, которые удовлетворяют двум изложенным условиям, относятся к одному классу, оставшиеся к другому.

Для проверки информативности сгенерированного набора признаков необходимо было выбрать один из алгоритмов классификации на два класса. Было принято решение использовать алгоритм, наиболее устойчивый к неинформативным, шумовым признакам, решающие леса, состоящие из 10 деревьев, реализованные в пакете Spider для MATLAB [27].

Информативность сгенерированных признаков может быть измерена как точность алгоритма на скользящем контроле. Этот выбор показателя информативности обусловлен тем, что модельная задача близка к задаче, которую ставят перед собой эксперты. Кроме того, ошибка на скользящем контроле является несмещенной оценкой качества алгоритма.

При фиксированных параметрах сегментации $C_{pair} = 20$, $C_1 = C_2 = 1$, алгоритмом была получена точность около 65 процентов на сбалансированной выборке. При соотношении числа объектов, отнесенных к классу "взорвавшихся" и "не взо-

рвавшихся" как в реальной жизни (разумеется, вспышки происходят непостоянно), достигаемая точность превосходит 85 процентов.

9 Автоматический подбор параметров сегментации

Нами используется схема вычисления информативных признаков в несколько этапов: сегментация, локализация активных областей, и уже после этого вычисление информативных признаков. При использовании многошаговых алгоритмов встает вопрос о том, как настраивать параметры на каждом из этапов: невозможно получить напрямую зависимость получаемых признаков от параметров унарных и парных потенциалов марковских случайных полей.

Обозначим процедуру сегментации за S, она зависит от параметров $\Theta = \{C_{pair}, C_1, C_2\}$. Процедуру локализации и подсчета признаков обозначим за L, оценку информативности за O_{cv}. Набор входных изображений обозначим за I_{train}, ответы о принадлежности объектов к одному из классов за Z_{train}. Таким образом, задача подбора параметров случайных марковских полей может быть записана следующим образом:

$$O_{cv}(L(S(I_{train}, \Theta)), Z_{train}) \to_{\Theta} \max$$

В явном виде решить эту задачу не представляется возможным, так как вид каждой из процедур сложно выписать в явном виде. Поэтому, для увеличения информативности генерируемых признаков мы перебрали несколько значений параметров потенциалов Θ. Таким образом мы автоматически подобрали параметры, оптимальные для генерации информативных признаков.

На рисунках 19 — 22 представлены результаты сегментации для разных значений параметров.

В ходе перебора был выявлен ряд фактов о влиянии каждого из параметров на точность результирующего алгоритма.

- При уменьшении параметра C_{pair} качество начинает немного падать, это связано с тем, что на сегментированном изображении появляется много шумовых областей.

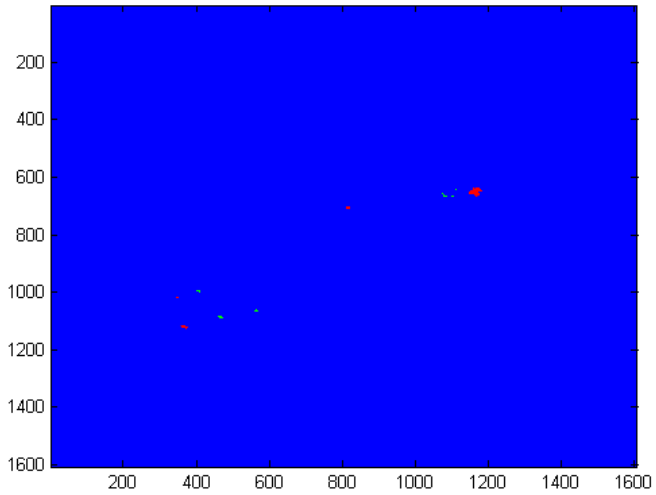

Рис. 19: Результаты сегментации при значениях параметров $C_{pair} = 20$, $C_1 = 0.2$, $C_2 = 0.05$

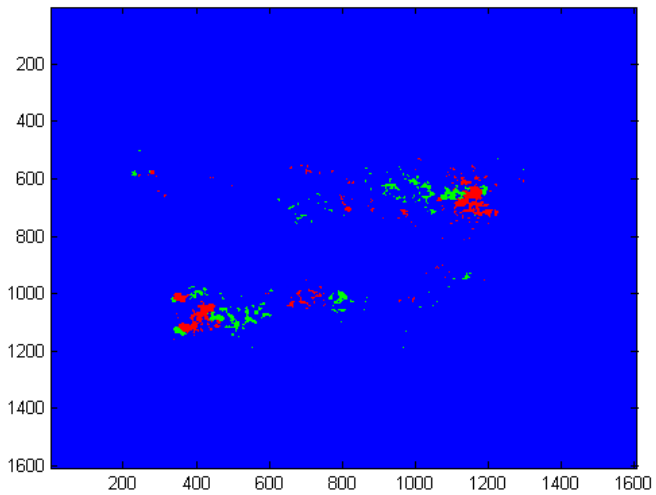

Рис. 20: Результаты сегментации при значениях параметров $C_{pair} = 20$, $C_1 = 0.2$, $C_2 = 0.1$

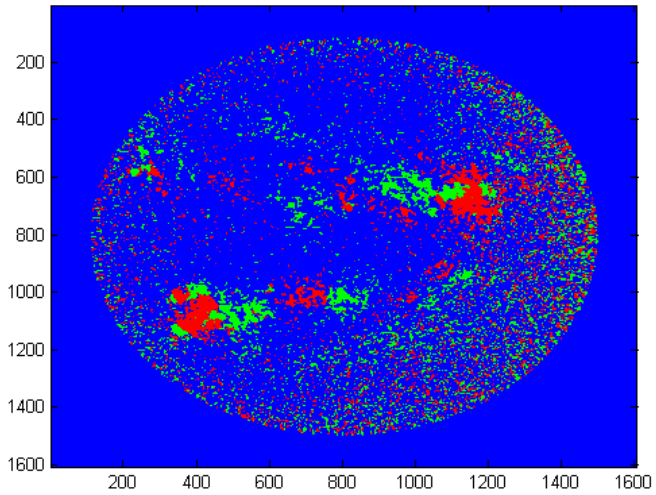

Рис. 21: Результаты сегментации при значениях параметров $C_{pair} = 20$, $C_1 = 0.2$, $C_2 = 0.5$

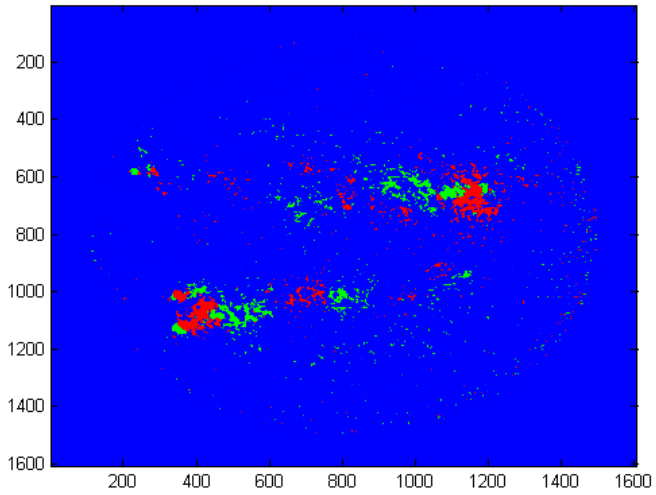

Рис. 22: Результаты сегментации при значениях параметров $C_{pair} = 1$, $C_1 = 1$, $C_2 = 1$

- Увеличение параметра C_{pair} более 20 не дает никакого прироста точности.
- Пропорциональное небольшое изменение параметров C_1 и C_2 практически не меняет вид сегментации, и, соответственно, не влияет на результирующую точность.
- Небольшое уменьшение параметра C_2 при фиксированном значении параметра C_1 приводит к уменьшению количества шумовых областей.

Кроме того, основным результатом работы по подбору параметров стал набор значений, на котором была получена точность, превосходящая точность на наборе параметров, фиксированным изначально "на глаз". Так, при значениях параметров: $C_{pair} = 20$, $C_1 = 0.2$, $C_2 = 0.3$ была получена точность около 67 процентов на модельной задаче по сравнению с точностью около 65 процентов при ручной подборке параметров экспертами.

10 Заключение

Была построена программная система, которая автоматически генерирует информативные признаки на основании последовательности магнитограммных снимков Солнца. Сгенерированные признаки использованы в последствии для задачи прогнозирования солнечной активности. Показанные методы генерации признаков на основании изображений могут быть успешно применены для других задач.

Результатами, в частности, являются:

- Разработана система автоматического подбора параметров марковских случайных полей.
- С помощью системы автоматического подбора параметров получена точность на модельной задаче выше, чем при подборе параметров экспертами.
- Выбран функционал, характеризующий активность области, построен алгоритм поиска активных областей на изображении.
- Разработан формализованный алгоритм нахождения "нейтральной линии" между областями с высоким положительным и отрицательным полями.

- Информативность сгенерированного набора признаков доказана на модельной задаче с помощью метода решающих лесов.

Список литературы

[1] *Чернышов, В.* Краткосрочное прогнозирование солнечной активности // Дипломная работа. Московский государственный университет имени М. В. Ломоносова, 2011.

[2] Официальный сайт Института Космических Исследований // http://www.iki.rssi.ru/

[3] *McIntosh P.* The Classification Of Sunspot Groups // Solar Physics, том 125, стр. 251-267, 1990.

[4] *Colak T., Qahwaji R.* Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images // Solar Physics, том 248, номер 2, стр. 277-296

[5] *Smith S., Howard R.* Magnetic Classification of Active Regions // International Astronomical Union. Symposium no. 35, стр. 33, 1968

[6] *Li R., Wang H., He H., Cui Y., Du Z.* Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting // Chinese Journal of Astronomy and Astrophysics том 7, номер 3, 2007

[7] *Falconer D., Moore R., Porter J.* Nutral-line magnetic shear and enhanced coronal heating in Solar active regions // The Astrophysical Journal, том 482 стр. 519-534, 1997

[8] *Tyan Y.* Polarity neutral lines on the solar surface and magnetic structures in the corona // Solar Physics, том 107, номер 2, стр. 247-262

[9] *Falconer D., Barghouty A., Khazanov I., Moore R.* A tool for empirical forecasting of major flares, coronal mass ejections, and solar particle events from a proxy of active-region free magnetic energy // Space Weather, 9, 2011

[10] *Blake A., Zisserman A.* Visual Reconstruction // The MIT Press Classics Series, 1987

[11] *Geman D., Geman S.* Parameter estimation for some Markov random fields // Pattern Analysis No. 11, 1983

[12] *Greig D., Porteous B., Seheult A.* Exact maximum a posteriori estimation for binary images // Journal of the Royal Statistical Society Series B, том 51, стр. 271–279, 1989

[13] *Wainwright M., Jaakkola T., Willsky A.* Map estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches // IEEE Transactions on Information Theory, выпуск 51(11), стр. 3697–3717, 2005

[14] *Kolmogorov V.* Convergent Tree-Reweighted Message Passing for Energy Minimization // IEEE Transactions on Pattern Analysis and Machine Intelligence, том 28(10), стр. 1568-1583, 2006

[15] *Boykov Y., Kolmogorov V.* An Experimental Comparison of Min-cut/Max-flow Algorithms for Energy Minimization in Vision // EMMCVPR 2001, стр. 359-374

[16] *Boykov Y., Veksler O., Zabih R.* Fast approximate energy minimization via graph cuts // Pattern Analysis and Machine Learning Intelligence 2001, стр. 1222-1239.

[17] *Scherrer P.* The solar oscillations investigation — Michelson Doppler imager // Solar Physics, том 162, стр. 129, 1995

[18] Официальный сайт программы по снимкам MDI // http://soi.stanford.edu/

[19] Definition of The Flexible Image Transport System (FITS) // FITS Standard v.3.0, 2008

[20] Официальный сайт обсерватории SOHO // http://sohowww.nascom.nasa.gov/

[21] *Vetrov D., Osokin A.* Submodular Decomposition Approach for Inference in Markov Random Fields // Интеллектуализация Обработки Информации, 2010

[22] *Michael J., Ghahramani Z., Jaakkola T., Saul L.* An introduction to variational methods for graphical models // Machine Learning том 37, стр. 183–233, 1999.

[23] *Woodford O., Rother C., Kolmogorov V.* A Global Perspective on MAP Inference for Low-Level Vision. // International Conference on Computer Vision, 2009

[24] *Kropotov D., Laptev D., Osokin A., Vetrov D.* Variational Segmentation Algorithms with Label Frequency Constraints // Pattern Recognition and Image Analysis, 2010

[25] *Bishom C.* Pattern Recognition and Machine Learning // Springer, 2006

[26] *Lampert C., Blaschko M., Hofmann T.* Beyond Sliding Windows: Object Localization by Efficient Subwindow Search // Computer Vision and Pattern Recognition, 2008

[27] Официальный сайт пакета Spider для MATLAB // http://people.kyb.tuebingen.mpg.de/spider/